

CANCER DIAGNOSIS USING SPECTRAL ANALYSIS AND INTERPRETATION

Vidan M.Fathi Ghoniem¹, Nahed Hussien Souloma², Abou Baker M. Youssef³, Yasser M. Kadah⁴

¹ Misr University for Science and Technology, 6th of October City, Egypt.

² Laser Institute for Enhanced Science, Cairo University, Giza, Egypt

^{3,4} Biomedical & Systems Department, Faculty of Engineering, Cairo University, Giza, Egypt

Abstract- Breast cancer comes as number one in ranking of malignancy tumors in the National Cancer Institute (NCI) of Egypt, Cairo University. Several changes occur during the transformation from normal to cancerous breast tissues. Such changes in molecular composition can be detected by Raman spectroscopy (RS). This study seeks to evaluate the ability of Raman spectra to distinguish breast tissues of different pathologies.

We used twenty two ex-vivo Raman samples of breast tissues, where histopathology was used as the gold standard. Based on these tissue samples, many computer algorithms were tried on these samples for the purpose of feature extraction and classification. Several spectral features contribute to the differences observed in normal and abnormal tissues. More elaborate techniques such as multivariate statistics methods were applied to precisely extract the discriminate features. These techniques include principal component analysis (PCA) and Fisher discriminate analyses (FDA) were used for data reduction and linear discrimination respectively. Independent component analysis (ICA) was also used to obtain the most independent data variables to remove any redundant information.

We also tried many pattern recognition approaches such as artificial neural network (ANN). It gave at a hand an easy-to-use diagnostic tool. We used back propagation and vector quantization networks. Other statistical techniques were also been attempted to classify the two different sets of data. The results of these techniques are encouraging and reveal the potential of optical biopsy in breast cancer diagnosis.

I. INTRODUCTION

Current detection methods for breast cancer include annual breast examination, biopsy, ultrasound, and magnetic resonance imagery (MRI). Mammography, a technique that investigates changes in tissue density, has several limitations. It diagnoses approximately five to fifteen percent of lesions that are non-palpable or are in well-known as 'difficult to diagnose' areas of the breast and about seventy to ninety percent of the lesions identified with mammography are ultimately deemed benign [1]. A technique called fine needle aspiration biopsy (FNAB) is proven to increase detection of malignancies in non-palpable lesions by up to thirty percent but is still insufficient because it disrupts the structure of the breast tissue and makes the results unreliable [2,3].

Unfortunately, cancer often goes undetected until it is well developed and responds to treatment poorly. Spectroscopic techniques are of great interest because they have the potential to identify cancer in its early stages. They are proved to be more sensitive, specific, and can be non-invasive [4].

Changes in cell and tissue structure associated with the disease are often apparent in tissue spectrum using Raman spectroscopy in particular, which may help detect the disease before it fully develops [5]. If the goal of early detection of primary or recurrent tumors can be achieved,

this may increase the likelihood of successful radical treatment and reduce complications [6]. Raman spectroscopy is a vibrational spectroscopic technique that originates from inelastic scattering of light by vibrating molecules. A Raman "spectrum" displays intensity as a function of frequency difference (the Raman shift) between the incident and scattered light. Thus, Raman spectroscopy provides detailed information about the bio-molecular composition of tissues, which might be used to distinguish between malignant tissues and normal; more specifically, non-neoplastic tissues.

The aim of this work is to develop reliable computer analysis algorithms to distinguish between non-neoplastic and cancerous breast tissue samples in vitro to assess the validity of optical biopsy for breast diagnosis using Fourier Transform Raman Spectroscopy (FTRS).

II. METHODOLOGY

2.1 Tissue Sample Preparation

Raman spectra were acquired from intact, ex vivo samples of human breast. The volume sampled was 1 mm³. These samples were acquired from the National Cancer Institute. The samples were used without any preparations; staining, or mounting, just fresh tissues (restricted within 7 hours). The samples were examined first by pathologic tests. So, we already knew which sample is normal and which is malignant.

2.2 Data Acquisition

The Raman spectra were acquired using FTRS by the help of the Laser Interaction with Matter Dept., National Institute of Laser Enhanced Science (NILES), Cairo University.

2.3 Data Analysis

The language used in this work is Matlab 7.0 Mathwork, Inc., from the Matlab tool boxes we used the signal processing tool box, the neural network tool box, the multivariate statistics tool box.

2.3.1 Preprocessing

The preprocessing and enhancement techniques are algorithms used to enhance the appearance of the spectrum such as base line restoration, normalization, maximization and others, which help with both qualitative and quantitative interpretation of spectra. Each spectrum was undergone many preprocessing steps to be suitable for the pattern recognition algorithms described below.

2.3.2 Pattern Recognition

A fundamental objective for pattern recognition is classification: given an input of some form we can analyze that input to provide a meaningful categorization of its data content. A pattern recognition system can be considered as a two stage device. The first stage is feature extraction. The second is classification.

2.3.2.1 Feature Extraction Methods

The region of interest has a wide wave-number range that may include a large number of significant intensity bands and hence we need to select only those have the highest discrimination power. The idea behind selecting these features is a trial to minimize the classification error while getting the most efficient discrimination. For every extracted set of features, we applied the t-test; which is a statistical measure for validating their significance in representing different sets of data and distinguishing between them.

First, the samples were examined visually. We then used four different methods to extract the most discriminate features that would help in better and more reliable classification.

Fisher discriminate analysis:

Dimensionality reduction is the aim of this method. We want to obtain discriminate features such that it maximizes the following criteria in (1):

$$J = ((\mu_1 - \mu_2)^2 / (\sigma_1 + \sigma_2)) \quad (1)$$

Where, μ_1 , σ_1 and μ_2 , σ_2 are the mean and variance of the two classes we are training; which represent the normal and tumor samples, respectively. This factor is measured for all features.

From the above equation, we can see that the Fisher factor, J has a higher value when the feature value differs greatly in the two classes and vice versa. Also the J value tends to zero when the difference between the means of the feature in the two classes tends to zero [8]. So, we can utilize the Fisher factor in extracting the most discriminate features. By applying T-test hypothesis on the output patterns we have got the following results shown in table (1).

Table (1), Ttest2 hypothesis applied on the fisher features

Positions	Significance value	Decision
F1 (3172.7)	0.00036	Y
F2 (3174.6)	0.00039	Y
F3 (3188.1)	0.000017	Y
F4 (3190)	0.0000225	Y
F5 (3192)	0.00003	Y
F6 (3381)	0.000019	Y
F7 (3382.9)	0.000018	Y

Principal Component Analysis (PCA):

In the automated methods of feature extraction, we used the principle component analysis to extract the most discriminate features from the training set. The principle component analysis is a method used in multivariate statistics to analyze the data set more obviously. In this method we try to estimate a set of orthogonal bases that upon which the data could be projected to minimize the effect of noise and to eliminate the redundant bases which have some degree of dependency on other bases. The principle components are linear combination of the original components. The most significant principle components are those having the highest variances. Mathematically, the principle components are estimated as follows:

Let the data to be analyzed are arranged in an $m \times n$ matrix X , where m is the number of measurement types and

n is the number of data trials. We have to estimate some matrix E that transforms X into a new data set Y . This represents a change of basis. The rows of E represent the set of new basis vectors that can be used to transform X into Y . These bases are to be orthonormal to reduce the redundancy. The eigenvectors of the covariance matrix of the original data set satisfies this condition. We can obtain up to m eigenvectors i.e. m orthonormal bases which will represent the principle components. To reduce the number of bases, we select the most principle components that are the eigenvectors having the largest eigenvalues. And these will represent the set of new bases. The calculation steps are as follows:

1. Estimate S_x , the covariance matrix of X .
2. Obtain the eigen values λ_i and the eigenvectors e_i of S_x , i from 1 to m .
3. Arrange e_i according the values of λ_i in a descending order.
4. Choose the first k vectors that have highest variances to represent the most principle components, i.e. the new bases.
5. Transform X into Y using e_i , i from 1 to k , as described above to get the new representation of the data.

By applying T-test hypothesis on the output patterns, we have got by applying inner product between the data and the three principal components of highest variances; we got the following results shown in table (2). Table (2), T-test applied on three features obtained by projecting the data on the new co-ordinates

PCs	Significance value	Decision
PC1	0.0018	Y
PC2	0.1331	N
PC3	0.3396	N

A plot of the first three columns of new data shows the data projected onto the first three principal components.

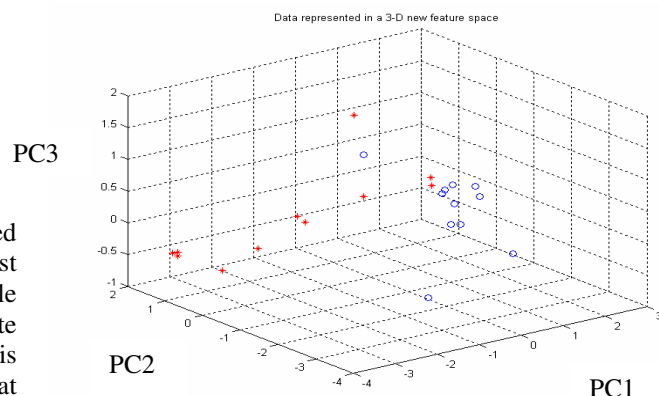


Fig .2. New data after projection on the PCs, represented in 3-D space (* for normal, o for tumour)

Independent component analysis (ICA):

It is a method for finding underlying factors or components from multivariate (multidimensional) statistical data. What distinguishes ICA from other methods is that it looks for components that are both statistically independent

and non Gaussian. ICA can be seen as an extension to principal component analysis and factor analysis. ICA is a much more powerful technique, however, capable of finding the underlying factors or sources when these classic methods fail completely.

In fact, ICA could be considered as non Gaussian factor analysis. In factor analysis it is often claimed that the factors are independent, but this is only partly true, because factor analysis assumes that the data has a Gaussian distribution. If the data is Gaussian, it is simple to find components that are independent, because for Gaussian data, uncorrelated components are always independent [9].

The first thing to note, that independence is a much stronger property than uncorrelatedness. Uncorrelatedness in itself is not enough to separate the components. PCA or factor analysis give components, that are uncorrelated but little more. This is the starting point of ICA. We want to find statistically independent components, in the general case where the data is non Gaussian.

The calculation steps are as follows:

1. Whiten the observed data; means that we linearly transform the observed data x by linearly multiplying it with some orthogonal matrix V , known as mixing matrix.
2. For whitened data z , we seek for a linear combination $w^T z$ that maximizes non-gaussianity, where w is known as separating matrix.
3. Estimating the independent components can be accomplished by finding the right linear combinations of the mixture variables.
4. Measuring the non-gaussianity using Kurtosis, where it is a classic measure for non-gaussianity for ICA estimation.

By applying T-test hypothesis on the output patterns, we have got by applying inner product between the data and the three independent components of highest variances; we got the following results shown in table (3).

Table (3), T-test applied on three features obtained by projecting the data on the new co-ordinates

ICs	Significance value	Decision
IC1	0.0017	Y
IC2	0.0052	Y
IC3	0.6816	N

Chemical Constituents-based Feature Extraction

The biological tissues are composed of chemical components (mainly proteins and lipids). Every component has a location (wave number) at which the intensity of scattering is maximal. So, in this study we observed the given spectra to specify the locations of different tissue constituents as suggested by the biochemistry group, Laser Interaction with Matter Dept., NILES, Cairo, University. We explored new features by using the intensity patterns corresponding to those locations, which represent the presence of different chemical components.

The data obtained correspond to the results achieved by many groups. Normal tissue spectra are dominated by the characteristic peaks of fatty acids, with Raman shifts of

1657, 1442 and 1300 cm^{-1} , whereas the lesions are dominated by structural protein modes at 1667, 1452, 1260, 890 and 820 cm^{-1} [1]. The main spectral differences of a typical FT-Raman spectra of normal, and malignant breast tissue at the interval of 600 to 1800 cm^{-1} were found in the bands of 1230 to 1295 cm^{-1} , 1440 to 1460 cm^{-1} and 1650 to 1680 cm^{-1} , assigned to the vibrational bands of the carbohydrate amide III, CH₂ bending-mode in proteins and lipids, and carbohydrate amide I, respectively. [10, 11].

By considering figure (3), the bands in the region between 2800- 3100 cm^{-1} , we found the CH₂ asymmetric and symmetric stretching modes at 2920 and 2851 cm^{-1} , respectively, are generally the strongest bands in the lipid spectra. The frequencies of these bands are conformation-sensitive. This is also the case for the bands due to the terminal CH₃ groups at 2956 cm^{-1} (asymmetric stretching) and 2873 cm^{-1} (symmetric stretching). The =C-H stretching bands due to unsaturated acyl chains are found at 3012 cm^{-1} [7]. In our results, we found the band in this region is higher in non-neoplastic tissue than in cancerous one which may attribute to the increased lipid content in non-neoplastic tissue than cancerous tissue.

In the region between 3080- 3500 cm^{-1} , there is a broad band which appears in both non-neoplastic and cancerous tissues, but it is more intense in cancerous tissue. We may attribute this to this region is dominated by the frequency of amide A and B, the nature of vibration is N-H stretching, in resonance with overtone (2 \times amide II) [7].

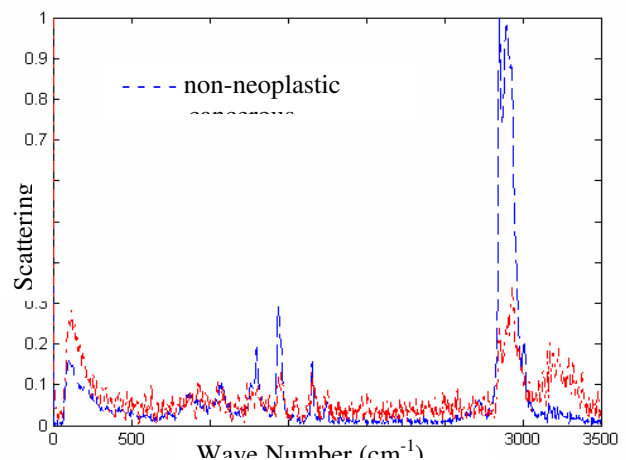


Fig.3. Raman spectra of breast tissue

2.3.2.2 Classification Techniques

The output stage is a classifier to assign a pattern to a certain class, or descriptor for the pattern. Generally speaking, the data set is usually divided into a training set and a test set. Some features are to be extracted from every input (training or test) to represent the pattern. The subject of pattern recognition is subdivided into: the statistical pattern recognition, and the neural network approach. [8].

Statistical Pattern Recognition

There exist many statistical classification methods such as the supervised and unsupervised learning methods. The unsupervised learning method, known as clustering, attempts to develop a representation for the given sample data. This method is used when the data set is unlabeled, such as k-means classifier. Using the unsupervised learning, subsets of these data may be formed into natural groupings or 'clusters', where each cluster most likely

corresponds to an underlying pattern class [8]. Though, we already had labeled data we used this classifier to make sure of the validity of the data sets. The non-parametric methods of classification such as the Euclidian distance classifier, and the nearest neighbor (KNN) classifier are of preference in our work.

Neural Networks

This strategy is attractive to the pattern recognition system designers, since the required amount of a priori knowledge and detailed knowledge of the internal system operation is minimal. Furthermore after training, we hope that the internal (neural) structure of the artificial implementation well self-organize to enable extrapolation when faced with new, yet similar, patterns, on the basis of experience with the training set [8]. In this work we used Kohonen network for unsupervised training and Back propagation (BP) and vector quantization (VQ) networks for supervised training.

III. RESULTS

A. Classification Results

Several techniques were applied to enhance the differentiation and classification of tissues for potential automated, clinical diagnosis. The simplest algorithm was based on visual inspection of the spectra and extracting spectral features. Then, using the FDA greatly enhances the performance of our classifiers as it offers the most discriminate features. A multivariate statistics was used; a powerful tool for dealing with different populations of tissue spectrum; to analyze Raman spectra. It was able to differentiate normal and malignant breast tissues. Furthermore, it helped in data reduction and removing the redundancy between data variables

We applied T-test hypothesis on all features in discriminating different pathologies. This test enabled us to test the significance of each feature and choose the most discriminate one before introducing it to the different classification algorithms we developed. Supervised pattern recognition used by ANN turned out to be helpful for tissue differentiation and identification and give at a hand an easy to use tool for rapid spectra analysis. The results of the network in classifying the 22 samples are shown in table (4.A) and (4.B).

Table (4.A), Back propagation results with all feature extraction techniques

Feature	All features	Fisher	PCA	ICA
Sensitivity %	100%	90.9%	90.9%	90.9%
Specificity %	72.7%	90.9%	81.8%	72.7%

Table (4.B), Learning vector quantization results with all feature extraction techniques

Feature	All features	Fisher	PCA	ICA
Sensitivity%	100%	100%	100%	100%
Specificity%	81.8%	100%	72.7%	81.8%

Supervised statistical pattern recognition used by Euclidean distance measurement and KNN classifier gave reliable results. The results we got in classifying the 22 breast samples is shown in table (5) and (6).

Table (5), Euclidean results with all feature extraction techniques

Feature	All features	Fisher	PCA	ICA
Sensitivity%	90.9%%	90.9%	100%	90.9%
Specificity%	72.7%	90.9%	72.7%	72.7%

Table (6), KNN results with all feature extraction techniques

Feature	All features	Fisher	PCA	ICA
Sensitivity%	81.8%	90.9%	81.8%	81.8%
Specificity%	81.8%	90.9%	72.7%	72.7%

Unsupervised pattern recognition used by Kohonen ANN and K-means classifier, gave the same results; 90.9% sensitivity and 72.7% specificity.

We explored new features by using the peaks corresponding to locations of different chemical tissue components; their central frequencies. By introducing these features to the classifiers, we developed the diagnosis according to the specific chemical/morphological make up of the tissues. The results of the classification algorithms are shown in tables (7 A-B).

Table (7 A), Lipid- based features

Classifier	ANN		Statistical	
Sensitivity%	81.8%	100%	100%	100%
Specificity%	72.7%	63.6%	72.7%	90.9%

Table (7 B), Protein-based features

Classifier	ANN		Statistical	
Sensitivity%	100%	100%	90.9%	90.9%
Specificity%	72.7%	72.7%	81.8%	81.8%

IV. DISCUSSION

Raman spectroscopy is a non destructive technique that uses specific excitation laser wavelength and laser power. So that tissue sample damage is avoided. It can identify tissue constituents and distinguish pathological changes. The abundance of diagnostic features in the tissue spectra clearly indicates the potential of this technique for clinical application. In comparing the features of the Raman spectra from breast tissues and their cancers, several similarities and differences were observed. Raman spectra of breast tissues both non-neoplastic and cancerous in the region 0-3500 cm^{-1} are shown in figure (3). The limitations

of this study are concerned with the small number of patients' samples. Our results show the potential of Raman spectroscopy in surveying the biochemical changes that accompany the development of neoplasia. However, larger histopathologic data are required to confirm these results.

V. CONCLUSION

In this study, a total of twenty two Raman spectra were acquired from intact, ex vivo samples of human breast. We developed a computer system to classify these breast tissue samples and distinguish between normal and cancerous ones. The algorithms we developed involve signal preprocessing as a primary stage. Secondly, pattern recognition, where the samples are examined visually to obtain the most discriminate features to be introduced to the classifiers to improve the classification process. We also, used FDA to obtain more discriminate features. In addition, we employed multivariate statistical analysis. In particular, PCA and ICA were used. After extracting these features, we applied T-test hypothesis on each feature to validate its significance before introducing it to different classifiers. The results we got are very reliable. We then, explored more features based on the chemical composition of the tissues; this helped in diagnosis according to the morphologic and chemical make up of breast tissues. From the results we got, we can observe that the classifiers attained best results with FDA features ; 100% sensitivity and 100% specificity with vector quantization network, and 90.9% sensitivity and 90.9% specificity with Euclidean and KNN classifiers. Finally, we conclude that the study shows that it is possible to measure Raman spectra in vitro and extract potentially diagnostically useful information.

REFERENCES

- [1] R. Manoharan: "Raman Spectroscopy and Fluorescence Photon Migration: for Breast Cancer Diagnosis and Imaging", *Photochemistry and Photobiology*, 67, 15-22(1998).
- [2] N.J. Kline and P.J. Treado: "Raman Chemical Imaging of Breast Tissue", *Journal of Raman Spectroscopy*, 28, 119-124(1977).
- [3] D.C.B. Redd: "Raman Spectroscopic Characterization of Human Breast Tissues: Implications for Breast Cancer Diagnosis", *Applied Spectroscopy*, 47, 787-791(1993).
- [4] G. Ullas: " Laser Raman Spectroscopy: Some Clinical Applications", *Current Science*, 77, 908-914(1999).
- [5] D. Pappas: "Raman Spectroscopy in Bioanalysis", *Talanta*, 51, 131-144(2000).
- [6] G. A. Wagnieres, W. M. Star, and B. C. Wilson: "In Vivo Fluorescence Spectroscopy and Imaging for Oncological Applications", *Photochemistry and Photobiology*, 68(5), 603-632(1998).
- [7] B. Stuart: "Biological Applications of Infrared Spectroscopy", *JOHN WILEY & SONS, INC.*, (1997).
- [8] R. J. Schalkoff: "Pattern Recognition: Statistical, Structural, and Neural Approaches", *JOHN WILEY & SONS, INC.*, (1992).
- [9] A. Hyvarinen, J. Karhunen, and E. Oja: "Independent Component Analysis", *JOHN WILEY & SONS, INC.*, (2001).
- [10] A. M. Jansen and R. R. Kortum: "Raman Spectroscopy for the Detection of Cancers and Precancers", *Journal of Biomedical Optics* 1(1), 31-70, (1996).
- [11] A. Martin, R. A. Bitar, Jr. Silveira, L., M. Zampieri, and M. M. Neto: "FT-Raman Spectroscopy Study for Human Breast Cancer Diagnosis", *Proceedings of SPIE*, 5141, Poster, (2003).